THE ROLE OF AUTOMATED SPEECH AND AUDIO ANALYSIS IN SEMANTIC MULTIMEDIA ANNOTATION

Franciska de Jong and Roeland Ordelman and Arjan van Hessen
University of Twente, Enschede, The Netherlands

***Abstract*— This paper overviews the various ways in which automatic speech and audio analysis can be deployed to enhance the semantic annotation of multimedia content, and as a consequence to improve the effectiveness of conceptual access tools. A number of techniques will be presented, including the alignment of text resources, large vocabulary speech recognition, key word spotting and speaker classification. The applicability of techniques will be discussed from a media crossing perspective. The added value will be illustrated by the description of two complementary demonstrators for browsing broadcast news archieves.**

I. INTRODUCTION

As is widely acknowledged, the exploitation of linguistic content in multimedia archives can boost the accessibility of multimedia archives enormously. Already in 1995, [3] demonstrated the use of subtitling information for retrieval of broadcast news videos, and in the context of TRECVID [11] the best performing video retrieval systems always exploit speech transcripts. Of course the added value of linguistic data is limited to video data containing textual and/or spoken content, or to video content with links to related textual documents, e.g. subtitles, generated transcripts etc. But when available, the use of linguistic content for the generation of time-coded indexes can help to bridge the semantic gap between media features and user needs. This will be explained in more detail for (a) (semi)-automatic annotation and (b) supporting conceptual querying of video content at fragment-level.

The semantic gap between user needs and content features is as old as the concept of archiving itself. The traditional approach towards the creation of indexes is to rely on manual annotation with controlled vocabulary index terms. With the emergence of digital archiving this approach is still widely in use and for many archiving institutes the creation of manually generated metadata is and will be an important part of the daily work. When the automation of metadata generation is considered, it is often seen as something that can enhance the existing process rather than replace it. The available metadata will therefore often be a combination of highly reliable and conceptually rich annotations, and (semi)automatically generated metadata. One of the challenges for search environments is to combine the various types of metadata and to exploit the added value of the combination. In this paper we will explain how available high level annotations for media archives can be exploited for improved automated generation of additional language-based annotations, and *vice versa*, how automatic content processing can help to generate ontological and thesaurial media annotations. For the content-based processing tasks the main focus will be on the various ways in which automatic speech and audio analysis can be deployed.

In the next sections we first explore some methods that deal with the exploitation of already available linguistic content in, or attached to multimedia databases. We carry the utilization of available collateral content further and introduce the concept of cross-media mining in section IV. Automatic audio indexing techniques are overviewed in the sections that follow: speech indexing in section V, speaker classification and emotion recognition in section VI. Finally, the added value of links that are automatically generated across media via high level annotation will be illustrated in section VII. This section will provide a description of two complementary demonstrators: one for on-line access to an archive of news broadcasts linked up to a newspaper archive, the other illustrating a crucial aspect in browsing multimedia databases, a technique known as *document clustering* applied in combination with topic detection.

II. EXPLOITING COLLATERAL TEXT

Depending on the resources available within an organization that administers a media collection, the amount of detail of the metadata and their characteristics may vary. Large national audiovi-

sual institutions such as Beeld&Geluid in The Netherlands[1], annotate at least titles, dates and short content descriptions (descriptive metadata). Many organizations with multimedia collections however, often do not have the resources to apply even some basic form of archiving.

To still allow the conceptual querying of video content, collateral textual resources that are closely related with the collection items can be exploited. A well known example of such a textual resource is subtitling information for the hearing-impaired (e.g., *CEEFAX* pages 888 in the UK) that is available for the majority of contemporary broadcast items, in any case for news programs. Subtitles contain a nearly complete transcription of the words spoken in the video items and provide an excellent information source for indexing. Usually, they can easily be linked to the video by using the time-codes that come with the subtitles. The Dutch news subtitles even provide topic boundaries that can be used for segmenting the news show into subdocuments. Textual sources that can play a similar role are teleprompter files: the texts read from screen by an anchor person (also referred to as auto-cues).

The time labels in these sources are crucial for the creation of a textual index into video. As in full text retrieval, where all words in a document can function as index terms and thus as a link to a document, the exploitation of collateral transcriptions for video will allow that all words spoken offer a link to the fragment in which they occur. And though full text retrieval is certainly not the ultimate solution to the semantic gap, natural language is inherently closer to the level of concepts that low-level image features. Also, there is a well-established range of generally applicable methods to turn text-based indexing into something that can be considered to support automatic semantic annotation, e.g., topic clustering and automatic classification. In principle these can all be applied in the domain of media archiving as well. Cf. [4] for an overview.

III. Time alignment

In the collateral text sources mentioned above, the available time-labels are not always fully reliable and can even be absent. In that case the text files will have to be synchronized. Examples of such text sources are minutes of meetings or written versions of lectures and speeches. This section will describe the approaches that we followed for the automatic generation of time-stamps for minutes in two pilot projects in the domain of e-Government. These minutes pertained to the so-called *Handelingen*, i.e. the meetings of the Dutch Parliament, and to city council meetings. Due to the difference in accuracy of the minutes, two different approaches had to be developed.

The minutes of the meetings of the Dutch Parliament are stenographic minutes that closely follow the discourse of the meeting, only correcting slips of the tongue and ungrammatical sentences. Given the close match with the actual speech, a relatively straightforward so-called forced alignment procedure could be used. Forced alignment is a technique commonly used in acoustic model training in automatic speech recognition (ASR). In order to be able to train phone models, words and phones in pre-segmented sentences are aligned to their exact location in the speech segments using an acoustic model[2]. Given a set of words from a sentence the acoustic model tries to find the most optimal distributions of these words given the audio signal on the basis of the sounds the words are composed of. When using alignment for indexing, pre-segmented sentences are evidently not available but as long as the text follows the speech well enough, the word alignment can be found by using relatively large windows of text.

The alignment procedure works well even if some words in the minutes are actually not in the speech signal. However, if the text to be aligned does not match the speech too well, as was the case with city council meetings, and if the text segments are too large, the alignment procedure will fail to find a proper alignment. In order to produce suitable segments, we used a two-pass strategy, similar as proposed in [8], incorporating the following steps as depicted in Figure 1:

1) a baseline large vocabulary speech recognition system[3] is used to generate a relatively inaccurate transcript of the speech with word-timing labels, referred to as hypothesis;
2) the hypothesis is aligned on the word level to the minutes using a dynamic programming algorithm;
3) the positions where the hypothesis and the minutes match (a match may be defined as three words in a row are correctly aligned), so called ‘anchors’ are placed.

[1]Beeld&Geluid:`http://www.beeldengeluid.nl/`

[2]In the first iteration usually an ’averaged’ bootstrap model is used. The alignment and the model should improve iteratively

[3]Optionally the speech recognition is somewhat adapted to the task for example by providing it with a vocabulary extracted from the minutes

4) using the word-timing labels provided by the speech recognition system, the anchors are used to generate suitable segments;
5) individual segments of audio and text are accurately synchronized using forced alignment;

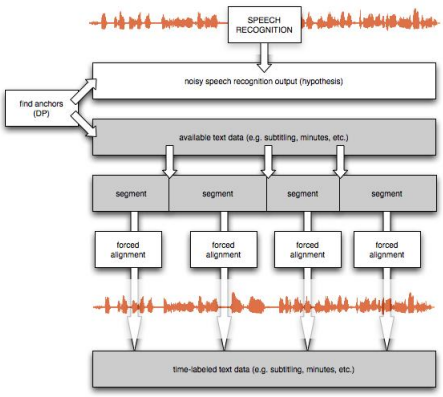

Fig. 1. Alignment procedure: synchronization of text and audio in a number of steps.

The described methods allow for the synchronization of audiovisual data to available linguistic content that approximates to a certain extent the speech in the source data and they enable the processing of conceptual queries of the audiovisual content with readily available tools.

IV. Cross-media mining

Ideally one would not only synchronize audiovisual material with content that approximates the speech in the data, but take even one step further and exploit *any* collateral textual resource, or even better: any kind of textual resource that is accessible, including open source titles and proprietary data (e.g., trusted webpages and newspaper articles). Another way of putting it: we propose to shift the focus from indexing individual multimedia documents to videomining in truly multimedia distributed databases. In the context of meetings for example, usually an agenda, documents on agenda topics and CVs of meeting participants can be obtained and added to the repository. Mining these resources can support information search because it yields annotations that offers the user not just access to a specific media type, but also different perspectives on the available data. An agenda could help to add structure that can for example be presented in a network representation, whereas CVs can be linked to annotations resulting from automatic speaker segmentation. In addition, both documents and CVs would allow for multi-source information extraction.

A typical example of what the cross-media perspective can yield in the broadcast news domain is the linking of newspaper articles with broadcast items and *vice versa*. Links can be established between two news objects which count is similar on the basis of the language models assigned to them via statistical analysis. Typically such language models are determined by the frequency of the linguistic units such as written or spoken words and their co-occurrences. The similarity between two documents can be decided for each pair of documents, but a more common approach is to pre-structure a document collection into clusters of documents with similar language models. Similarity of language models predicts similarity of topic, and therefore this technique is know as topic clustering. [4]

In addition to linking documents with a similar topic profile, which can be supportive in a browser environment, also the available semantic annotation for documents with similar profiles can be exchanged and exploited for conceptual search. If a newspaper article has been manually classified as belonging to e.g., economy or foreign politics, a broadcast item with a similar language model can be classified with these conceptual labels as well.

In section VII below, we describe a cross-media news browser demonstrator that incorporates this functionality. For the linking of audiovisual data with textual resources that are not directly related to the speech, such as documents on agenda topics in the context of meetings, or newspaper data in the broadcast news domain, we have to step up the use of speech recognition compared to the speech recognition deployed in the alignment procedures described in section III. In the more elaborate alignment procedure, an initial hypothesis is generated by a large vocabulary speech recognition system. As this hypothesis is only needed for finding useful segments, we do not really care about the performance of the system as long as it is able to provide us with 'anchors'. However, the relevance of speech recognition performance increases when textual resources suitable for alignment with audiovisual data are *not* available. In the next section, the application of speech recognition technology as the *primary* source for generating a textual representation of audiovisual documents

[4]The functionality commonly known as *topic detection and tracking* (TDT) for dynamic news streams has been built upon it and plays a central role in the evaluation series for TDT organized by DARPA.

that can be linked to other linguistic content, is described.

V. Speech indexing

Recent years have shown that large vocabulary speech recognition can successfully be deployed for creating multimedia annotations allowing the conceptual querying of video content and the synchronization to any kind of textual resource that is accessible, including other full-text annotations of audiovisual material. This is especially the case in the broadcast news domain which is very general and makes data collection for training a speech recognition system relatively easy. For the broadcast news domain, speech transcripts approximate the quality of manual transcripts, at least for several languages. Spoken document retrieval in the American-English broadcast news (BN) domain was even declared "solved" with the NIST-sponsored TREC SDR track in 2000 [5]. It should be noted however that in other domains than broadcast news and for other languages, a similar recognition performance is usually harder to obtain due to a lack of domain-specific training data, and in addition to a large variability in audio quality, speech characteristics and topics being addressed.

A. Vocabulary adaptation

One of the main topics in large vocabulary speech recognition in the context of broadcast news indexing is vocabulary adaptation. Given the huge quantities of available training data for the broadcast news domain, the acoustic models and language models can usually be trained adequately and in addition, various acoustic adaptation procedures (e.g., bandwidth/gender dependent models, speaker adaptive training) can be applied to boost ASR performance. However, the language model and vocabulary are usually created using a fixed, and with respect to broadcast news often outdated training corpus. Vocabulary words are normally selected on the basis of their frequency of occurrence in the data. This procedure is in contrast with the linguistic properties in broadcast news that continuously change over time: names of places and people are unexpectedly introduced, people that are frequent in the news during a period of time may disappear after a while, jargon may suddenly become public property and new words are invented. In addition, some words that are highly probable to occur in one period of the year and highly improbable in another period. Vocabulary selection based on frequencies in outdated text material will not be able to capture these dynamics.

In the context of retrieval it is highly relevant that the speech recognition system adapts to this linguistic variations in order to reduce the number of out-of-vocabulary (OOV) words and with that, the number of OOV query words (QOV). The latter are words that appear in a user's query and also occurred in the audiovisual document but –as they were OOV– could not be recognized correctly. OOV's damage retrieval performance in two ways: firstly, given a query with a QOV word, the QOV word leads to a word *miss* in searching. Secondly, its replacement potentially induces a *false alarm* for other queries.

Although document expansion and query expansion techniques may be deployed to compensate for QOV words ([6], [14]), tackling OOVs in an earlier stage is favorable, especially in the context of cross-media mining as particularly names may be important keys. For tuning up the speech recognition vocabulary in the broadcast news domain for new words up-to-date training text material is required. Ideally, one can dispose of a daily feed of newspapers but as such a corpus will often not be available, contemporary data can also be collected via the Web [1] or by capturing subtitling information from news programs. A number of selection methods have been proposed that make use of parallel corpora for vocabulary adaptation, such as using a small look-back time window to select new words based on frequencies [2], using word history information in a large parallel corpus [10] and a vectorial-based method combining several adaptation corpora [1] .

B. Word spotting

Another way of dealing with OOV words is using a word-spotting approach as an alternative or auxiliary tool. Word-spotting is sometimes combined with an approach based on phone transcriptions or phone lattices. Especially when the mismatch between speech recognition vocabulary and domain vocabulary is hard to model and tends to produce many out-of-vocabulary words, having word spotting functionality available as an ad-hoc tool for searching either the audio directly or a phone or phone lattice representation of the document (cf. [7]) can be profitable. A typical example of a deploying word spotting approach in combination with a full text transcription approach would be the following strategy to recover names that were misrecognised: (i) the initial speech recognition transcript is used to find related collateral text data; (ii) named entity detection in the collateral data source provides relevant named en-

tities given the document topic; (iii) the occurrence (and timings) of these named entities in the source data are recovered using a word spotting approach.

VI. Speaker Classification

There is more information in the speech than the words alone. Speaker characteristics can be extracted from the speech (speaker's voice, word usage, syntax) as well and may serve as an additional level of information, for example to add structure for browsing (speaker segmentation and identification) or to extract features that could not be accessed using traditional views on the data. Automatic speaker classification can especially be beneficial for spoken documents in cultural heritage collections. Such spoken word archives receive attention from professional information analysts from different fields. Historians for example, may be interested both in the exact words that were spoken, but also in the speaker's profile. The latter may partly be reconstructed using the identification of speaker characteristics such as accent, age, gender, speaking behavior and even emotion and cognitive state.

VII. Demonstrators

A number of techniques described above have been implemented in (currently) two separate demonstrators that illustrate how the concept of cross-media news browsing for a multifaceted multimedia archieve can be realized.

A. Cross-media news browser

The first demonstrator that we refer to as the cross-media news browser, was initially a demonstrator for on-line access to an archive of Dutch news broadcasts (NOS 8 uur Journaal). It shows how either available collateral data sources (subtitling information for the hearing-impaired) or full-text speech recognition transcripts can be used as linguistic content for the generation of time-coded indexes for searching within news shows. Although the subtitling information in itself would already be enough to enable access, speech recognition transcripts are generated as well for demonstration purposes. The subtitling information is captured using a teletext capturing card and synchronized with the video using a manually determined offset value. The speech recognition system consists of decision-tree state-clustered acoustic models trained on approximately 20 hours of speech from the Spoken Dutch Corpus [9], a vocabulary of 65K words extracted from a newspaper collection and a 3-gram language model trained on some 300M words of newspaper text data. Currently, the speech recognition system is static; it does not update the vocabulary and language model, nor does it perform any acoustic adaptation schemes. The incorporation of such procedures is scheduled for a new version of the demonstrator.

As the subtitling information provides information on topic boundaries, we can use real topic boundaries for the segmentation of the news show into 'subdocuments'. In the speech recognition case, we pretend that we do not have these boundaries and segment the news show on the basis of acoustic information such as speaker changes, speech/non-speech occurrences and silences. As segmentation based on acoustic information does not necessarily takes place on topic boundaries, segments belonging to the same topic should be clustered afterwards in order to generate topic based subdocuments. This is however not implemented yet.

In order to demonstrate the added value of links that are automatically generated across media via high level annotation, we linked up the linguistic annotations of the news items (either based on subtitles or ASR) to an up-to-date database of Dutch newspaper articles that we can use for demonstration purposes by courtesy of PCM publishers[5]. The newspaper articles are being indexed. The links from the news topics to the newspapers articles are generated by using the stopped video annotations as a query for searching the newspaper database.

B. Novalist news browser

Next to the broadcast news browser that primarily demonstrates the added value of automatic linguistic annotation of audiovisual content, at TNO a news browser for heterogenous media archives has been developed which is called Novalist. The functionality of this browser can be regarded as complementary for the news browser described above. It aims to facilitate the work of information analysts in the following way: (i) related news stories are clustered to create dossiers, sometimes also called 'threads', (ii) dossiers resulting from clustering are analysed and annotated with several types of metadata, and (iii) a browsing screen provides multiple views on the dossiers and their metadata.

The corpus disclosed by the Novalist demonstrator system consists of a collection of news

[5]PCM publishers is one of the largest publishers in the Dutch language region: `http://www.pcmuitgevers.nl/`. For reasons of IPR the public demonstrator does not contain the links to these articles.

items published by a number of major Dutch newspapers and magazines, web crawls, a video corpus of several news magazines and a video archive with all 2001 news broadcasts of *NOS 8 uur Journaal*. Here, the autocue files for the video archive function as collateral text. Transcripts of broadcast audio generated with automatic speech recognition (ASR) can also be incorporated. The entire collection currently consists of some 160,000 individual news items from 21 different sources.

Novalist demonstrates a crucial aspect in browsing multimedia databases, a technique known as *document clustering* applied in combination with topic detection. The system has to deal with dynamic information, about which no full prior knowledge is available. There is no fixed number of target topics and events types. The system must both discover new events as the incoming stories are processed, and associate incoming stories with the event-based story clusters already created. Clustering is done incrementally: for a new incoming story, the system has to decide instantaneously to which topic cluster the story belongs. Since the clustering algorithms are unsupervised, no training data is needed.

Via document clustering, structure is generated in news streams, while the annotations can be applied as filters: search for relevant items need not to apply on analyzed data but can be limited to relevant subsets of the collection. Novalist supports the fast identification of relevant dossiers during browsing. Dossiers are visualized in a compact overview window with links to a time axis. Additional functionality could consist of the automatic generation of links to related sources, both internal and external. For a detailed explanation of the concept of topic detection and the similarity concept applied in the language modeling approach that is underlying Novalist, and for an overview of the performance evaluation of some components, cf. [12], [4].

Novalist demonstrates that multiple document abstractions effectively mediate different levels of granularity. The analysis can be performed independently of end-user queries. Due to the emphasis on content preprocessing it can support an entire chain of users: content portals that select subsets of news according to filters to serve their users, professional information analysts that link the portal content to their own repositories, and nomadic news consumers. The broadcast news browser and the Novalist browser show how semantic multimedia annotation can succesfully be exploited.

ACKNOWLEDGEMENT

This work was partly supported by the Dutch bsik-programme MultimediaN (www.multimedian.nl), and the EU projects AMI (IST-FP6-506811) and MESH (IST-FP6-027685).

REFERENCES

[1] A. Allauzen and J.L. Gauvain. Diachronic Vocabulary Adaptation for Broadcast News Transcription. In *InterSpeech*, Lisbon, September 2005.

[2] C. Auzanne, J.S. Garofolo, J.G. Fiscus, and W.M Fisher. Automatic Language Model Adaptation for Spoken Document Retrieval. In *Proceedings of RIAO 2000, Content-Based Multimedia Information Access*, pages 132–141, 2000.

[3] M. G. Brown, J.T. Foote, G.J.F. Jones, K. Sparck Jones, and S. J. Young. Automatic Content-based Retrieval of Broadcast News. In *Proceedings of the third ACM international conference on Multimedia*, pages 35–43, San Francisco, November 1995. ACM Press.

[4] F.M.G. de Jong and W. Kraaij. Content Reduction for Cross-media Browsing. In H. Saggion and J.-L. Minel, editors, *RANLP workshop 'Crossing Barriers in Text Summarization Reserach*, pages 64–69, Borovets, Bulgaria, 2005.

[5] J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference*, pages 107–129, Washington, 2000.

[6] P. Jourlin, S.E. Johnson, K. Spärck Jones, and P.C. Woodland. General Query Expansion Techniques for Spoken Document Retrieval. In *Proc. ESCA Workshop on Extracting Information from Spoken Audio*, pages 8–13, Cambridge, UK, 1999.

[7] W. Kraaij, J. van Gent, R. Ekkelenkamp, and D. van Leeuwen. Phoneme based spoken document retrieval. In *Proceedings of the fourteenth Twente Workshop on Language Technology TWLT-14*, pages 141–153, University of Twente, 1998.

[8] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, 1998.

[9] N. Oostdijk. The Spoken Dutch Corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Second International Conference on Language Resources and Evaluation*, volume II, pages 887–894, 2000.

[10] R.J.F. Ordelman. *Dutch Speech Recognition in Multimedia Information Retrieval*. Phd thesis, University of Twente, Enschede, October 2003. publisher: Taaluitgeverij Neslia Paniculata publisherlocation: Enschede, ISSN: 1381-3617; No 03-56, ISBN: 90-75296-08-8, Numberofpages: 268.

[11] A.F Smeaton, W. Kraaij, and P. Over. Trecvid - an overview. In *Proceedings of TRECVID 2003*, USA, 2003. NIST.

[12] Martijn Spitters and Wessel Kraaij. Unsupervised clustering in multilingual news streams. In *Proceedings of the LREC 2002 workshop: Event Modelling for Multilingual Document Linking*, pages 42–46, 2002.

[13] Khiet P. Truong and David A. van Leeuwen. Automatic detection of laughter. In *InterSpeech*, pages 485–488, Lisbon, September 2005.

[14] P.C. Woodland, S.E. Johnson, P. Jourlin, and K. Spärck Jones. Effects of Out of Vocabulary Words in Spoken Document Retrieval. In *2000 ACM SIGIR Conference*, pages 372–374, Athens Greece, 2000.